le cnam

Expliquer ou prédire?
Données massives et nouveaux défis

Gilbert Saporta
CEDRIC- CNAM,
292 rue Saint Martin, F-75003 Paris

http://cedric.cnam.fr/~saporta

Plan

1. Les deux cultures
2. Parcimonie et complexité
3. Validation empirique
4. Paradigmes et paradoxes
5. Interpréter les modèles
6. Big Data: faux espoirs et défis
7. Comprendre pour mieux prédire

1. Les deux cultures

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

- The generative modelling culture
  - seeks to develop stochastic models which **fits** the data, and then make inferences about the data-generating mechanism based on the structure of those models. Implicit (…) is the notion that there is a **true model** generating the data, and often a truly `best' way to analyze the data.
- The predictive modelling culture
  - is silent about the underlying mechanism generating the data, and allows for many different predictive algorithms, preferring to discuss only accuracy of prediction made by different algorithm on various datasets. **Machine Learning** is identified by Breiman as the epicenter of the Predictive Modeling culture.

From Donoho, 2015

- Conception standard (modèles pour comprendre)
  - Fournir une certaine compréhension des données et de leur mécanisme générateur à travers une représentation parcimonieuse .
  - Un modèle doit être simple et ses paramètres interprétables pour le spécialiste : élasticité, odds-ratio, etc.
- En « Big Data Analytics » (modèles pour prédire)
  - Pour de nouvelles observations: généralisation
  - Les modèles ne sont que des algorithmes

Cf GS, compstat 2008

4 | Le Monde Mercredi 29 janvier 2014 | SCIENCE & MÉDECINE | ÉVÉNEMENT

« Big data »
Trois défis pour les maths

INFORMATIQUE

L'analyse de grandes masses de données pour en tirer des informations pertinentes est un domaine en pleine expansion. Les « data scientists » doivent imaginer de nouveaux algorithmes pour maîtriser les volumes, la vitesse et la variabilité de ce déluge numérique

- modèles classiques inadaptés
  - Tout est significatif!
    - si $n=10^6$ un coefficient de corrélation égal à 0,002 est significativement différent de 0 mais bien inutile
    - Modéles usuels rejetés
    - Intervalles de confiance de longueur nulle

Même formule: $y = f(x;\theta)+\varepsilon$

- **Modélisation explicative**
  - Théorie sous-jacente
  - Ensemble restreint de modèles
  - Qualité d'ajustement: prédire le passé
  - Erreur: bruit blanc
- **Modélisation prédictive**
  - Modèles issus des données
  - Modèles algorithmiques
  - Prédire de nouvelles données: prédire l'avenir
  - Erreur: à minimiser

- Du côté des économistes:

Symposium: Big Data

Big Data: New Tricks for Econometrics (pp. 3-28)
Hal R. Varian

Abstract/Tools | Full-text Article (Complimentary) | Download Data Set (2.62

AMERICAN ECONOMIC ASSOCIATION

AEA American Economic Association

AEA | Journals | Annual Meeting | EconLi

Journal of Economic Perspectives: Vol. 28 No. 2 (Spring 2014)

- "Data manipulation tools and techniques developed for small datasets will become increasingly inadequate to deal with new problems. Researchers in machine learning have developed ways to deal with large datasets and economists interested in dealing with such data would be well advised to invest in learning these techniques."

2. Parcimonie et complexité

- Le rasoir d'Ockham
  - *pluralitas non est ponenda sine necessitate*
  - Un principe scientifique : éviter des hypothèses inutiles

Guillaume d'Ockham (1285 – 1349), dit le « docteur invincible » philosophe, théologien et logicien franciscain.
Etudes à Oxford, puis Paris.
Accusé d'hérésie, convoqué pour s'expliquer à Avignon, excommunié pour avoir fui à Munich à la cour de Louis IV de Bavière. Meurt vraisemblablement de l'épidémie de peste noire. Réhabilité par Innocent VI en 1359

A inspiré le personnage du moine franciscain Guillaume de Baskerville dans le « Nom de la rose » d'Umberto Eco.
Premier jour, vêpres : « il ne faut pas multiplier les explications et les causes sans qu'on en ait une stricte nécessité. »

- AIC, BIC et autres vraisemblances pénalisées sont souvent considérées comme des versions modernes du rasoir d'Ockham

  $$AIC = -2\ln(L) + 2K$$

  $$BIC = -2\ln(L) + K\ln(n)$$

- Une similarité trompeuse : **AIC et BIC issus de théories différentes**
  - AIC : approximation de la divergence de Kullback-Leibler entre la vraie distribution f et le meilleur choix dans une famille paramétrée
  - BIC : choix bayesien parmi des modèles paramétriques a priori équiprobables
- **Illogique de les utiliser simultanément**

Comparaison AIC BIC

- L'AIC est un critère prédictif tandis que le BIC est un critère explicatif.
- Si *n* tend vers l'infini la probabilité que le BIC choisisse le **vrai** modèle tend vers 1, ce qui est faux pour l'AIC.
- Pour *n* fini: résultats contradictoires. BIC ne choisit pas toujours le vrai modèle: il a tendance à choisir des modèles trop simples en raison de sa plus forte pénalisation

AIC BIC réalistes?

- Vraisemblance pas toujours calculable.
- Nombre de paramétres non plus: ridge, PLS, arbres, etc.
- « **Vrai** » modèle?

  “Essentially, all models are wrong, but some are useful ”
  (G.Box,1987)

* Box, G.E.P. and Draper, N.R.:
Empirical Model-Building and Response Surfaces, p. 424, Wiley, 1987

- « Occam’s Razor, long admired, is usually interpreted to mean that simpler is better. Unfortunately in prediction, accuracy and simplicity (interpretability) are in conflict » Breiman, 2011

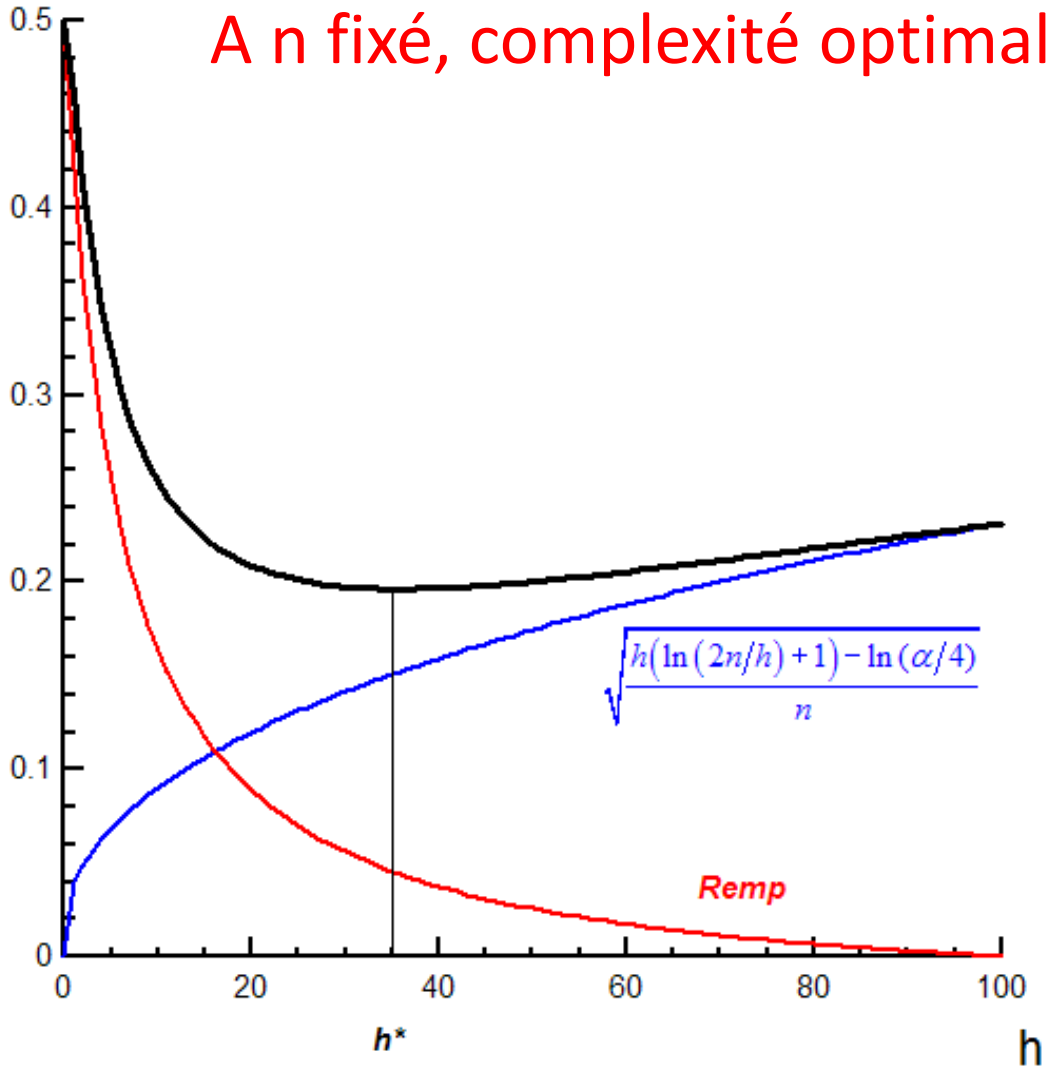
- Pénalisation en contradiction avec la théorie de l'apprentissage
- L'inégalité de Vapnik

$$R < R_{\text{emp}} + \sqrt{\frac{h\left(\ln\left(2n/h\right)+1\right)-\ln\left(\alpha/4\right)}{n}}$$

dépend de n/h d'où ces conséquences :

- On améliore la capacité prédictive si la complexité h croit mais moins vite que n
- On peut augmenter la complexité du modèle avec n

A n fixé, complexité optimale h*



- Les meta-modèles ou méthodes d'ensemble
  - Bagging, boosting, **random forests** améliorent les algorithmes élémentaires
  - Idem pour le **stacking** (Wolpert, Breiman) qui combine linéairement les prévisions de modèles de toutes sortes: linéaires, arbres, ppv, réseaux de neurones etc.

$$\hat{f}_1(\mathbf{x}),\ \hat{f}_2(\mathbf{x}),\ldots,\ \hat{f}_m(\mathbf{x})$$

$$\hat{y} = \sum_{j=1}^{m} w_j \hat{f}_j(\mathbf{x})$$

    - Avec des poids positifs de somme 1: version fréquentiste du Bayesian Model Averaging

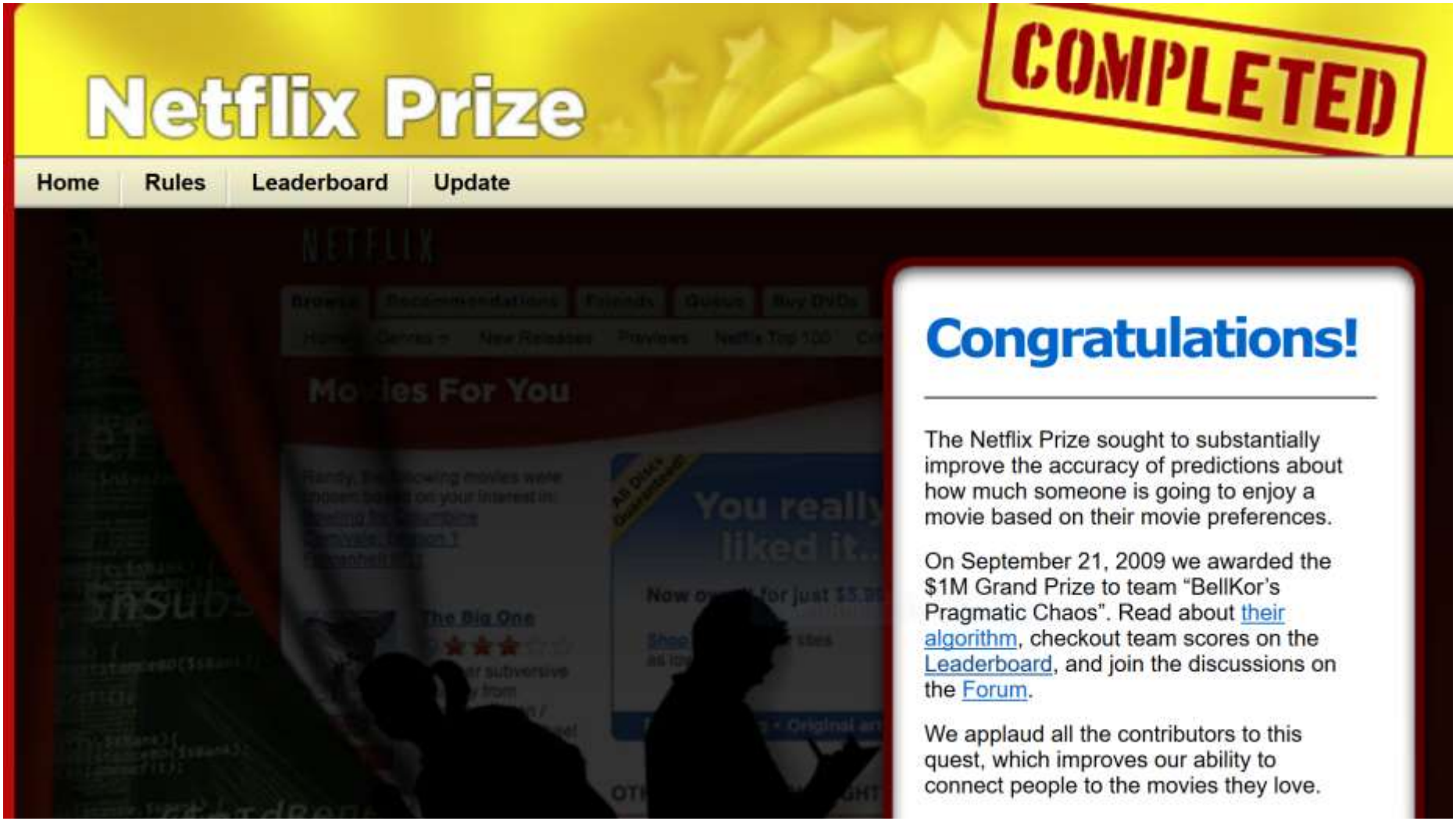
- Première idée: poids obtenus par mco:

$$\min \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} w_j \hat{f}_j(\mathbf{x}) \right)^2$$

  - favorise les modèles les plus complexes: surapprentissage
- Solution: utiliser les valeurs prédites par leave one out

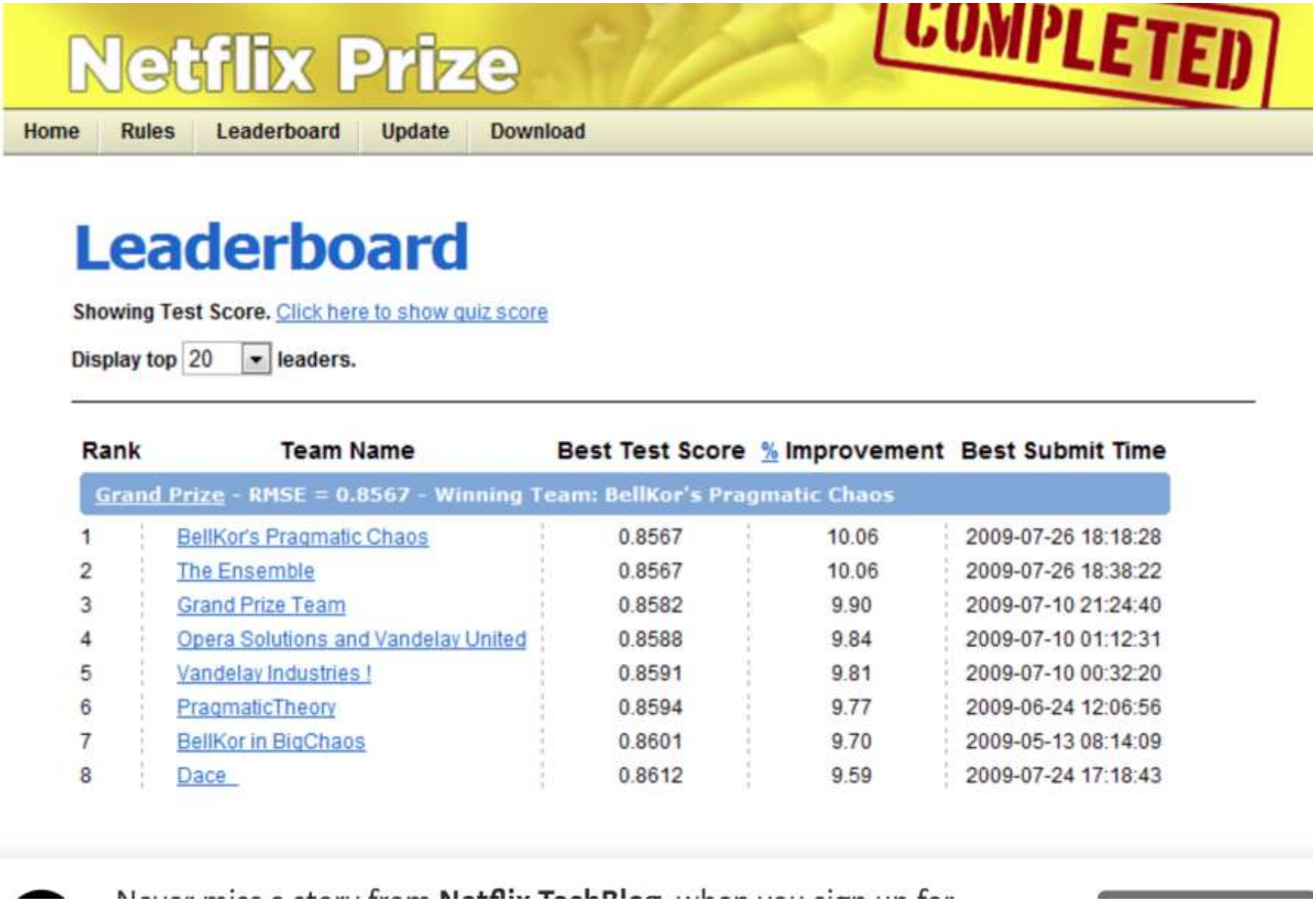
$$\min \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} w_j \hat{f}_j^{-i}(\mathbf{x}) \right)^2$$

- Améliorations : (Noçairi et al., 2016)
  - Régression PLS ou autre méthode régularisée car les m prévisions sont très corrélées
- Empiriquement le stacking surpasse sur de nombreux cas le BMA avec des calculs bien plus simples (Clarke, 2003)



We offered $1 million to whoever improved the accuracy of our existing system called *Cinematch* by 10%. The race was on to beat our RMSE of 0.9525 with the finish line of reducing it to 0.8572 or less

- The Netflix dataset contains more than 100 million datestamped movie ratings performed by anonymous Netflix customers between Dec 31, 1999 and Dec 31, 2005. This dataset gives ratings about m = 480 189 users and n = 17 770 movies
- The contest was designed in a training-test set format. A hold-out set of about 4.2 million ratings was created consisting of the last nine movies rated by each user (or fewer if a user had not rated at least 18 movies over the entire period).The remaining data made up the training set.
- We offered $1 million to whoever improved the accuracy of our existing system called *Cinematch* by 10%. The race was on to beat our RMSE of 0.9525 with the finish line of reducing it to 0.8572 or less


Netflix Prize
COMPLETED
Home
Rules
Leaderboard
Update
Download
Leaderboard
Showing Test Score. Click here to show quiz score
Display top 20 leaders.
Rank
Team Name
Best Test Score
% Improvement
Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos
1 BellKor's Pragmatic Chaos 0.8567 10.06 2009-07-26 18:18:28
2 The Ensemble 0.8567 10.06 2009-07-26 18:38:22
3 Grand Prize Team 0.8582 9.90 2009-07-10 21:24:40
4 Opera Solutions and Vandelay United 0.8588 9.84 2009-07-10 01:12:31
5 Vandelay Industries ! 0.8591 9.81 2009-07-10 00:32:20
6 PragmaticTheory 0.8594 9.77 2009-06-24 12:06:56
7 BellKor in BigChaos 0.8601 9.70 2009-05-13 08:14:09
8 Dace 0.8612 9.59 2009-07-24 17:18:43

- 44014 valid submissions from 5169 different teams
- *BellKor's Pragmatic Chaos team*. A blend of hundreds of different models
- *The Ensemble Team* . Blend of 24 predictions

  Same Test RMSE : 0.8567 (10.06%)
- Bellkor's Pragmatic Chaos defeated The Ensemble by submitting just 20 minutes earlier!

Mais Netflix n'a pas implémenté la méthode victorieuse:

We evaluated some of the new methods offline but the additional accuracy gains that we measured did not seem to justify the engineering effort needed to bring them into a production environment.

http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html

Electronic Journal of Applied Statistical Analysis
Vol. 09, Issue 02, October 2016, 340-361
DOI: 10.1285/i20705948v9n2p340

Improving stacking methodology for combining classifiers: applications to cosmetic industry

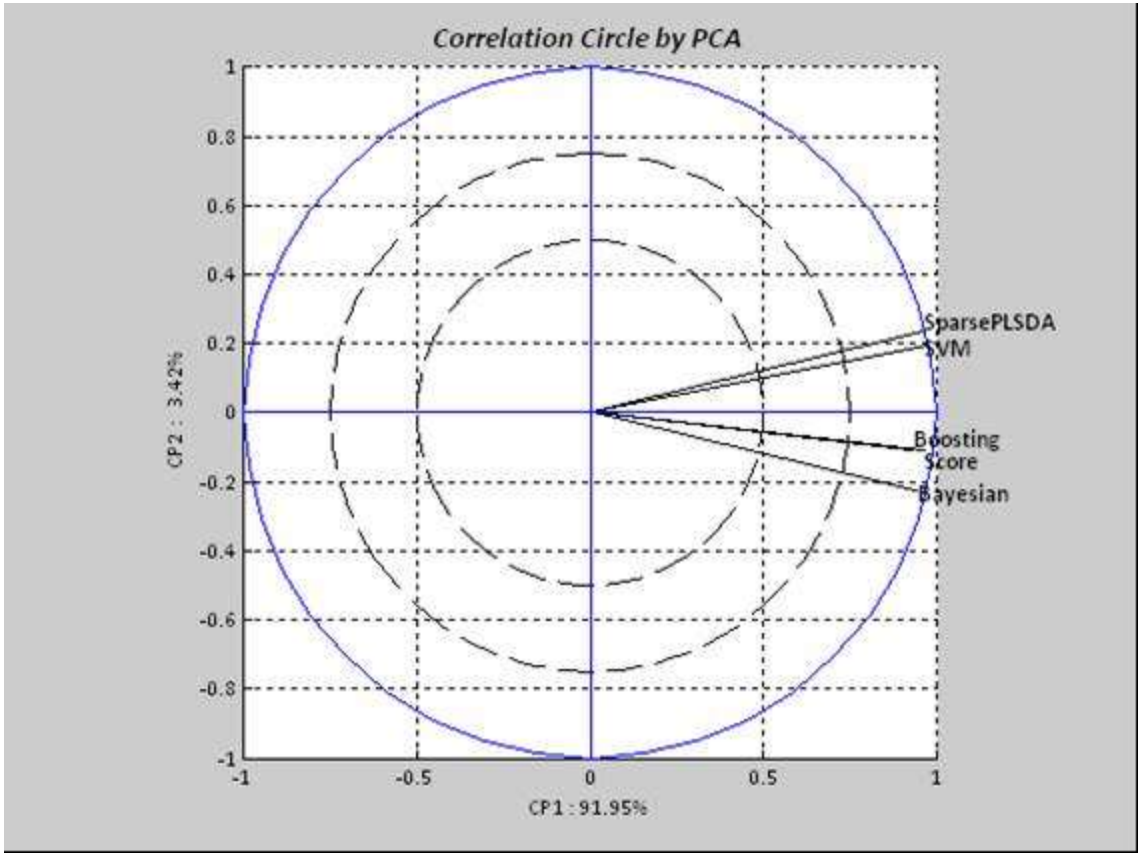
Noçairi H.*[a], Gomes C.[a], Thomas M.[a], and Saporta G.[b]

[a] *L'Oréal Recherche,1 avenue Eugène Schueller, BP22, 93601 Aulnay sous bois, France*
[b] *CEDRIC, CNAM , 292 rue Saint Martin, 75141 Paris cedex 03, France*

Statutory context and motivation:

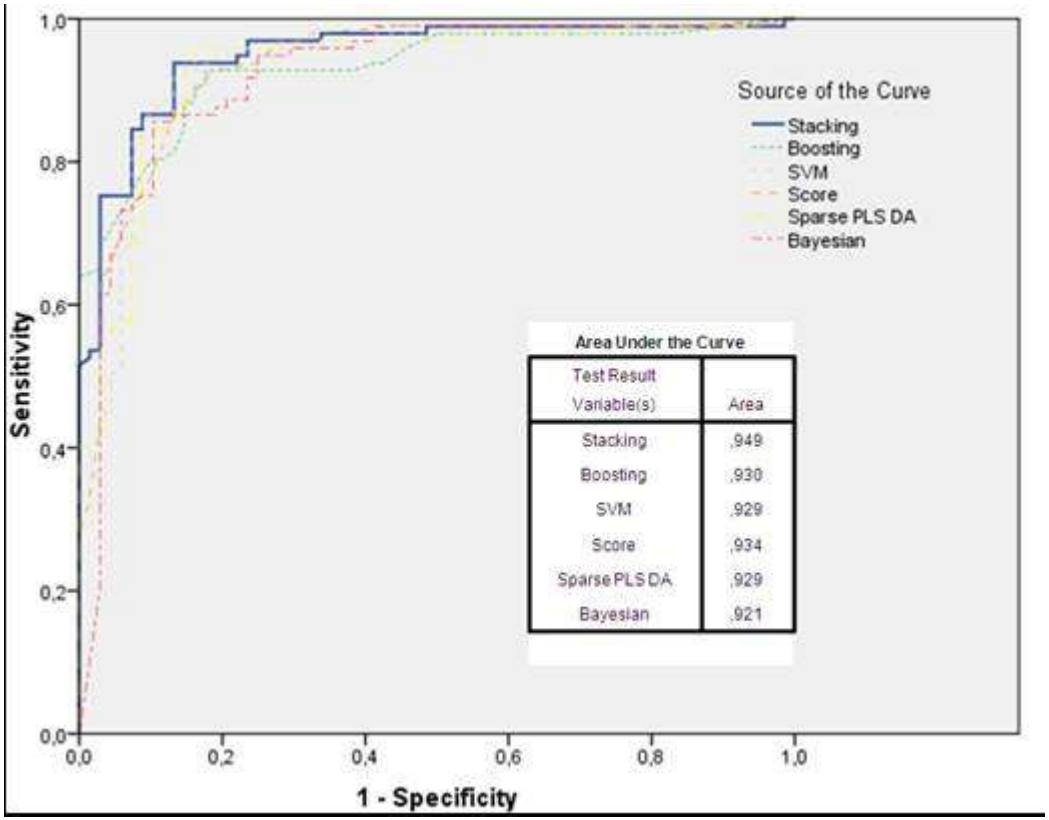
- According to the 7th Amendment of the European Cosmetic Directive concerning the stop of animal experimentation , L'OREAL developed several prediction models for various end-points such as the skin irritation, ocular irritation, and sensitization.
- Data set : 165 chemicals characterized by 35 variables; results from *in silico* predictions, *in vitro* test, assays as well as physicochemical experimental or calculated parameters.
- Binary response: Sensitizer/No-Sensitizer

– Five models give correlated predictions



Stacking with logistic PLS outperforms the 5 single models

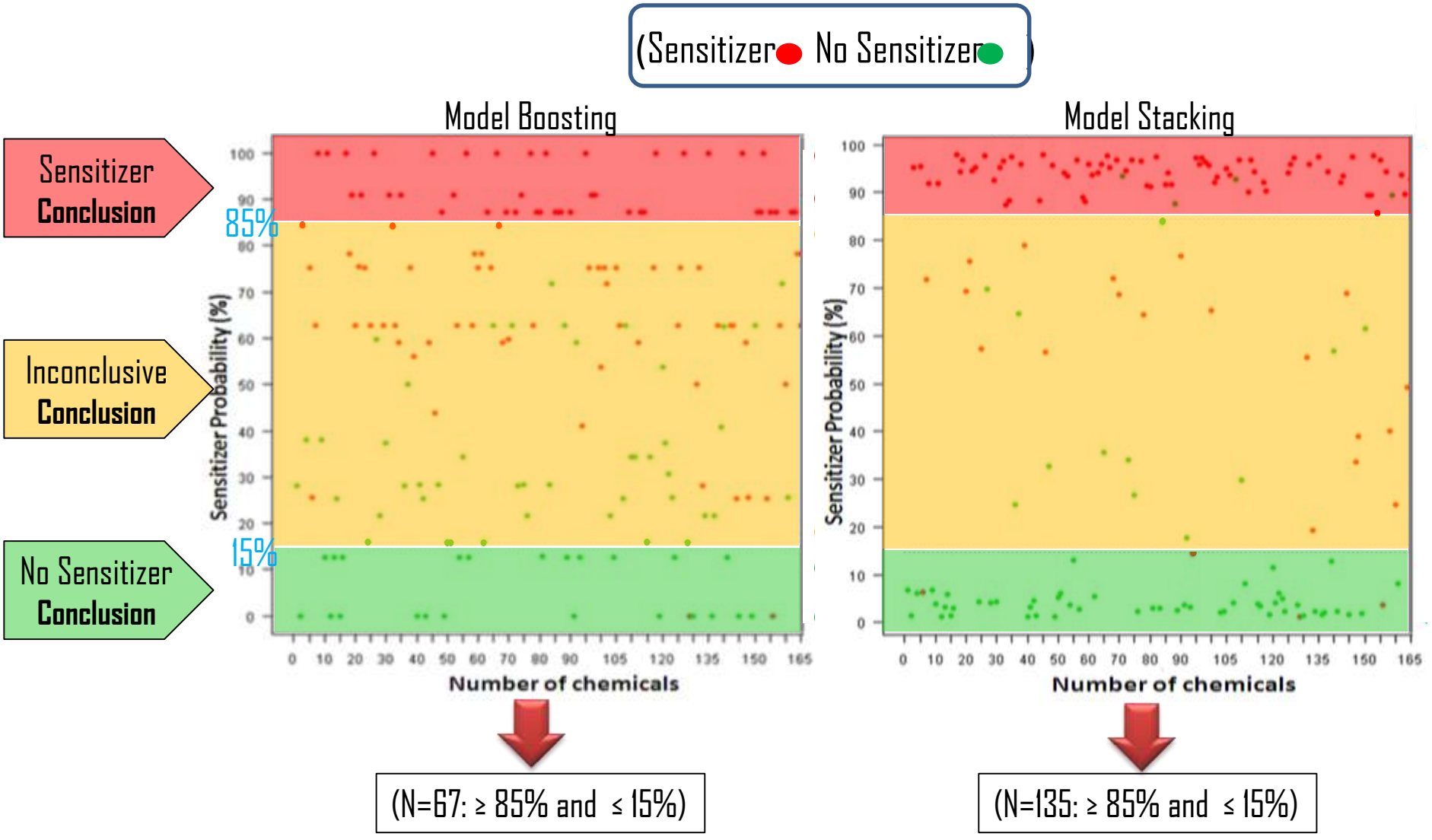
Stacking weights	
Model	w_j
Boosting	1.656
Score	2.281
Sparse PLS	1.311
SVM	1.609
Naive Bayes	1.188



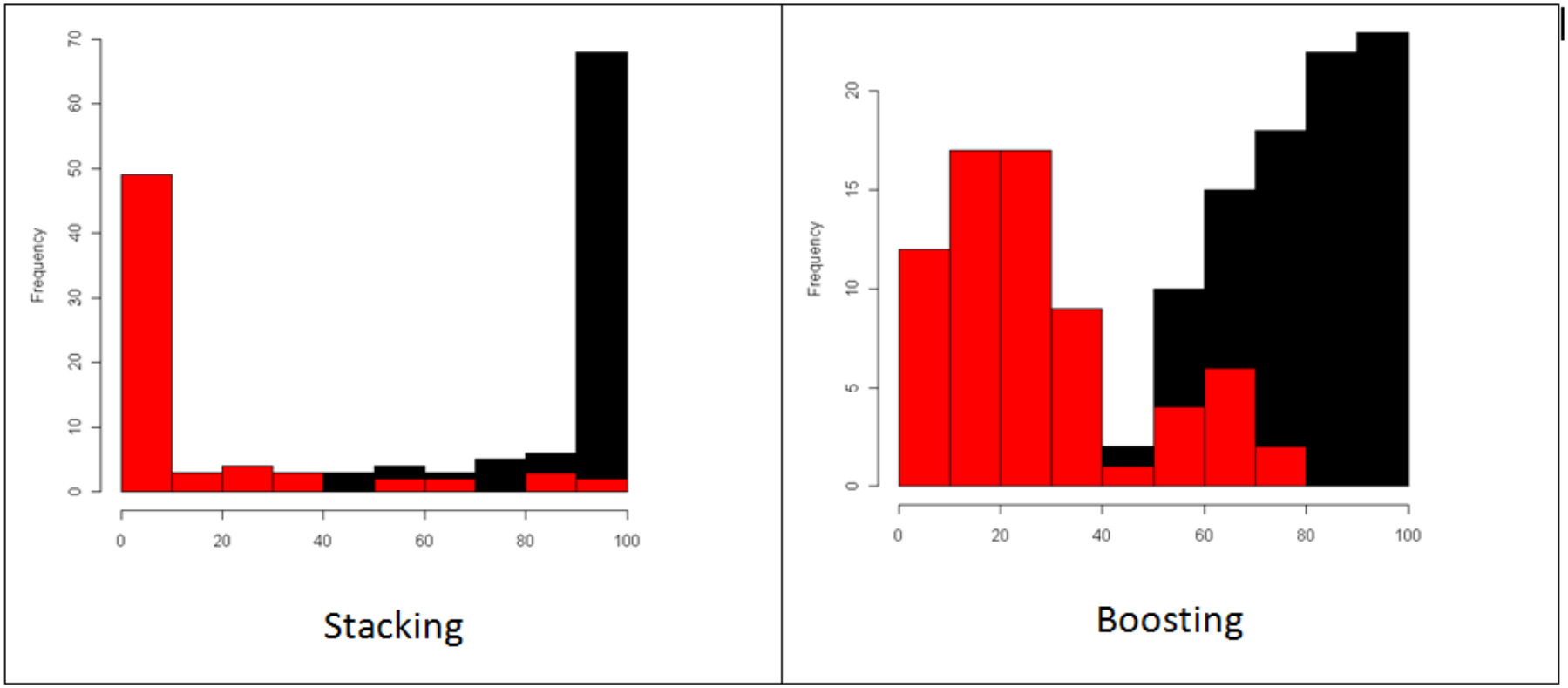
Decision rule

- chemicals with a probability >=85% are predicted Danger,
- chemicals with a probability <=15% are predicted Non Danger
- chemicals with a probability between those two thresholds are inconclusive

Stacking allows decision for more chemicals



Danger probabilities:
stacking provides a better bimodality



3. Validation

- Nécessité de marier Machine Learning et statistique
  - Un bon modèle est celui qui prédit bien
  - Différence entre ajustement et prévision
  - Contrôler le risque de surapprentissage
  - Ensembles d'apprentissage et de validation

Une démarche avec 3 échantillons pour choisir entre plusieurs familles de modèles:

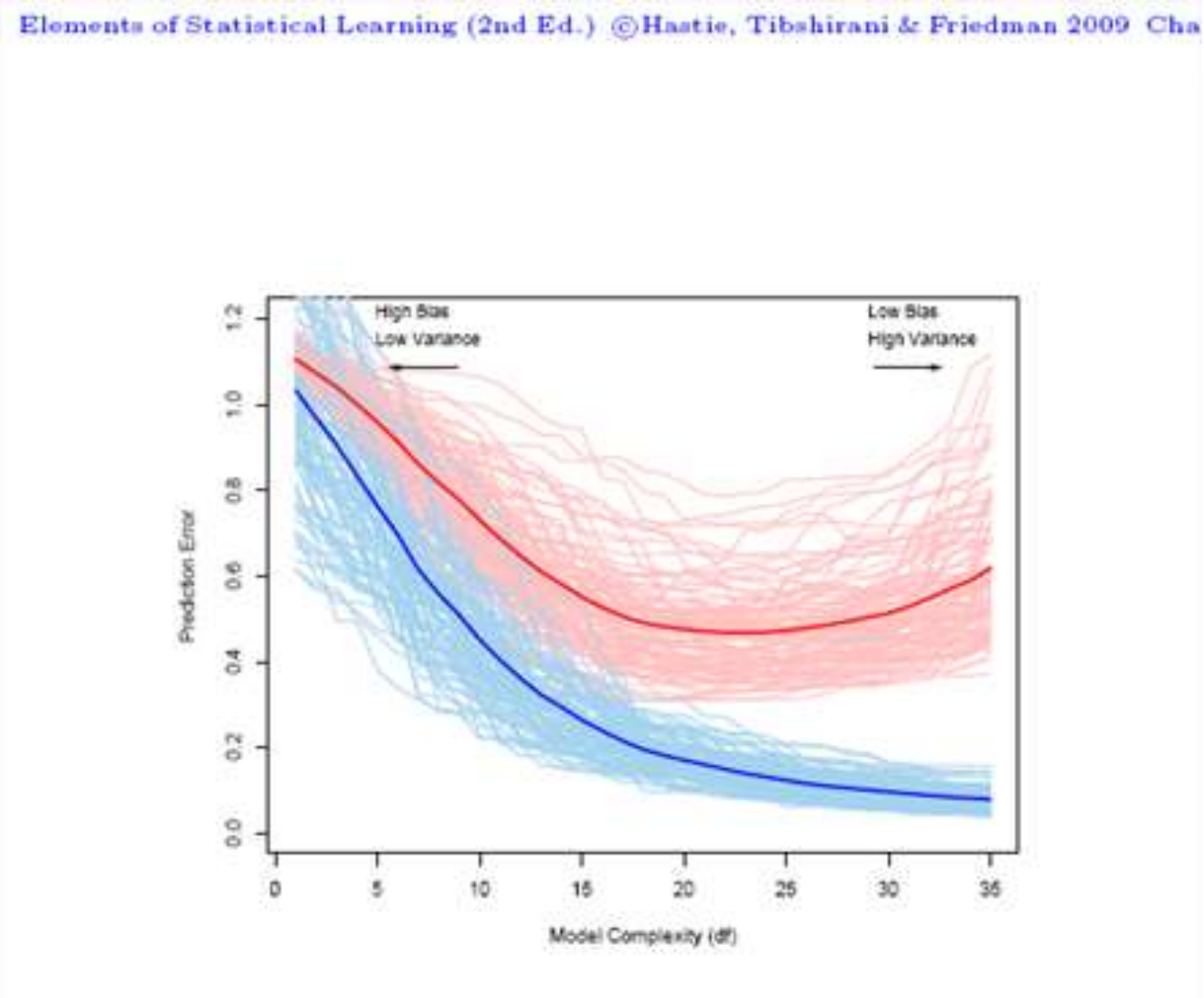
- Apprentissage: pour estimer les paramètres des modèles
- Test : pour choisir le meilleur modèle
- Validation : pour estimer la performance sur des données futures
  - **Estimer les paramètres ≠ estimer la performance**
  - Réestimation du modèle final**: avec toutes les données disponibles**

- Précurseurs:
  - Paul Horst (1903-1999)

    « *the usefulness of a prediction procedure is not established when it is found to predict adequately on the original sample; the necessary next step must be its application to at least a second group. Only if it predicts adequately on subsequent samples can the value of the procedure be regarded as established*» 1941
  - Leave one out : Lachenbruch et Mickey, 1968
  - Validation croisée: Stone, 1974

- **Elémentaire?**
  - Pas si sur…
  - Voir publications en économetrie, épidémiologie, .. prédictions rarement validées sur des données « hold-out » (sauf en prévision de séries temporelles)

- Séparer (une fois) les données en apprentissage, test et validation ne suffit pas



4. Paradigmes et paradoxes

- Comprendre sans prédire
  - Un « bon » modèle qui s'ajuste bien peut fournir des prévisions médiocres au niveau individuel (eg épidémiologie)
- Prédire sans comprendre
  - Des modèles ininterprétables (eg *deep learning*) peuvent donner de bonnes prévisions (ciblage marketing, …)

Le paradigme de la boîte noire

- modèle génératif y=f(x)+ ε . On cherche une fonction qui approxime en un certain sens le comportement de la boîte noire.
- Deux conceptions très différentes :
  - soit on cherche à approximer la vraie fonction f,
  - soit on cherche à obtenir des prévisions de y aussi précises que possible.

– Modern statistical thinking makes a clear distinction between the statistical model and the world. The actual mechanisms underlying the data are considered unknown. The statistical models do not need to reproduce these mechanisms to emulate the observable data (Breiman, 2001).

– Better models are sometimes obtained by deliberately avoiding to reproduce the true mechanisms (Vapnik, 2006).

- Garder (ou pas) des variables significatives ou non?
  - A researcher might choose to retain a causal covariate which has a strong theoretical justification *even if is statistically insignificant*.
  - statistical significance plays a minor or no role in assessing predictive performance. In fact, it is sometimes the case that removing inputs with small coefficients, *even if they are statistically significant*, results in improved prediction accuracy

(Shmueli, 2010)

- Le modèle « vrai » ne prédit pas toujours mieux
  - the underspecified linear regression model that leaves out q predictors has a lower EPE when the following inequality holds:

$$q\sigma^2 > \beta_2' X_2' (I - H_1) X_2 \beta_2.$$

  - when the data are very noisy (large σ);
  - when the true absolute values of the left-out parameters (in our example β_2) are small;
  - when the predictors are highly correlated; and
  - when the sample size is small or the range of left-out variables is small.

(Shmueli, 2010)

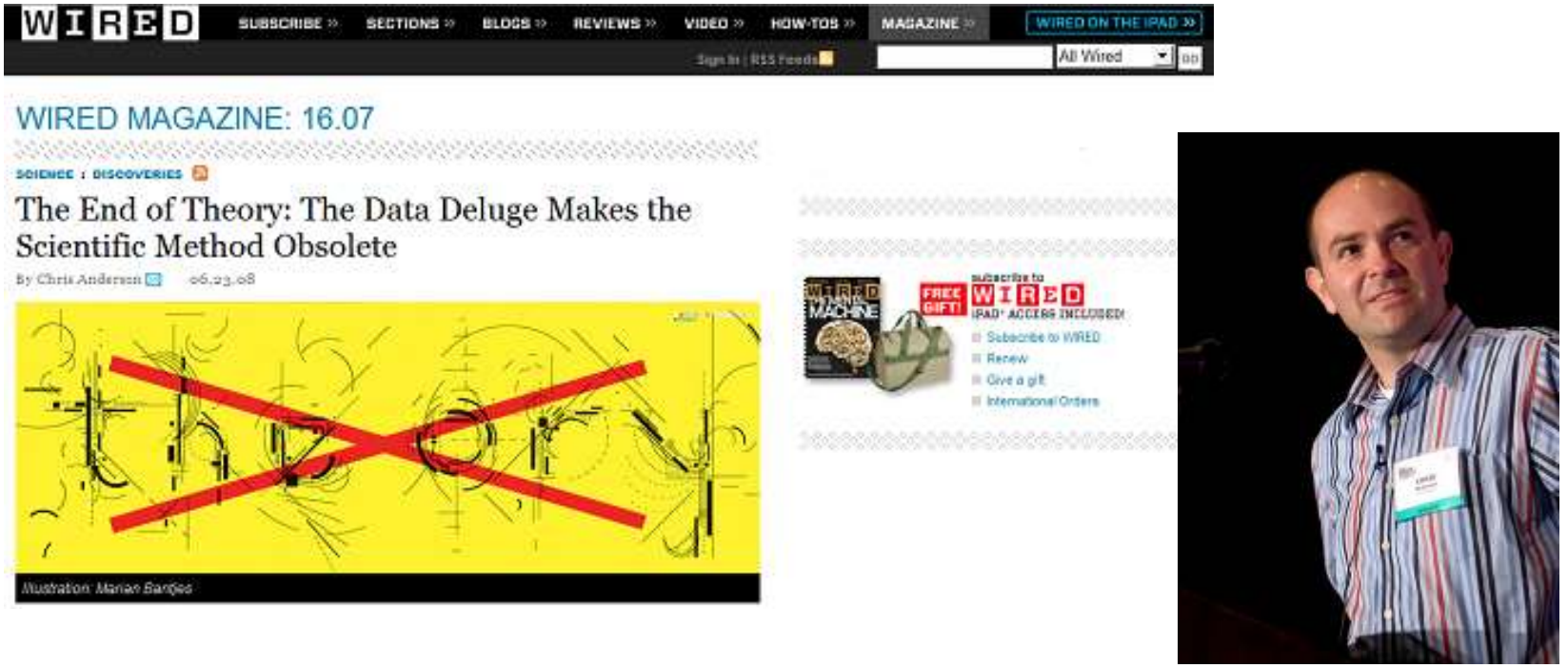
5. Interpréter les modèles

- On croit souvent que les modèles simples (régression linéaire, logistique) s'interprètent aisément
- C'est loin d'être le cas!
- Sauf dans des dispositifs orthogonaux, la valeur des paramètres ne reflète que rarement l'importance des variables

- Plus de 11 méthodes pour quantifier l'importance des variables en régression linéaire! (Grömping, 2015, Wallard, 2015)
  - Eg Shapley value: un sous-ensemble de prédicteurs vu comme une coalition en théorie des jeux
- Encore des idées de Breiman:
  - « A variable might be considered important if deleting it seriously affects prediction accuracy »
  - permutation aléatoire des valeurs « shuffling »
  - Utilisées dans les random forests mais applicable pour tout modèle « *model agnostic* »

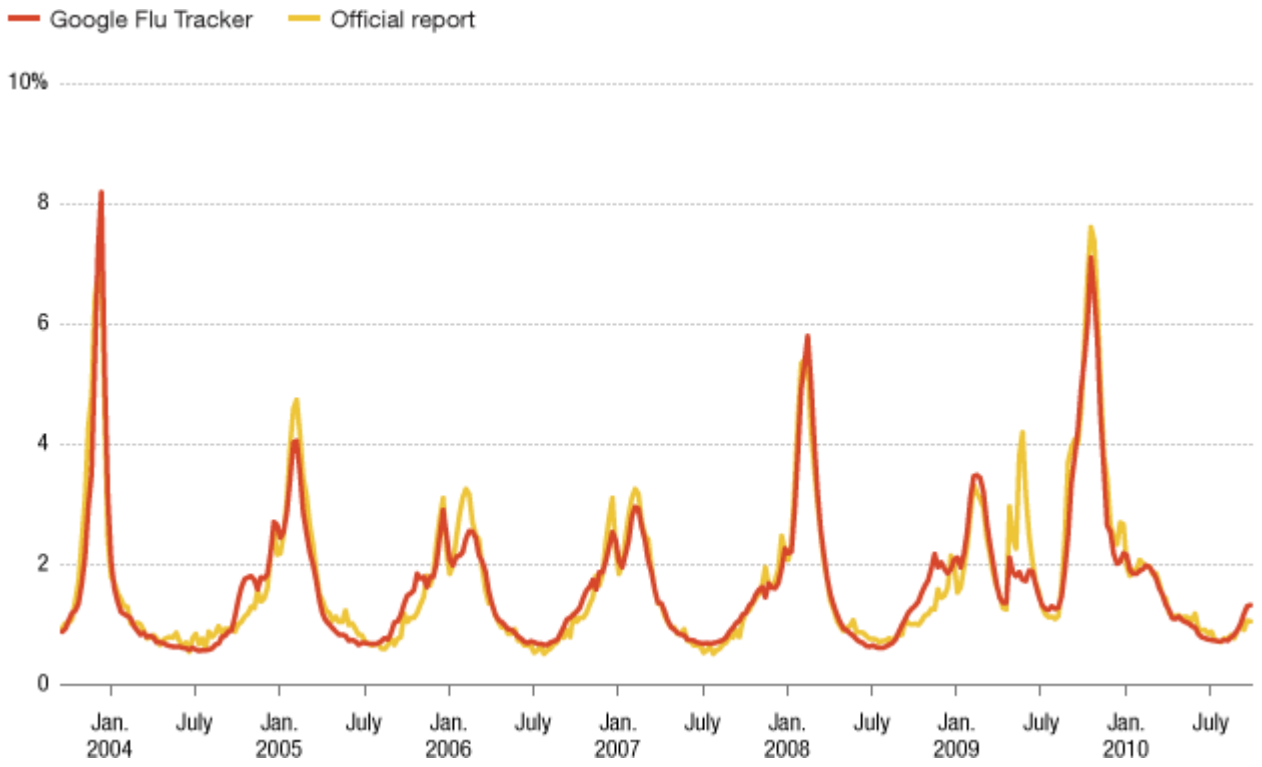
- Mais :
  - « toutes choses égales par ailleurs» souvent impossible
  - Faire varier un prédicteur (intervention) peut impliquer des variations des autres d'où un effet complexe
  - Nécessité de schémas causaux

6. Big Data: faux espoirs et défis



Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.

- Google FluTrends



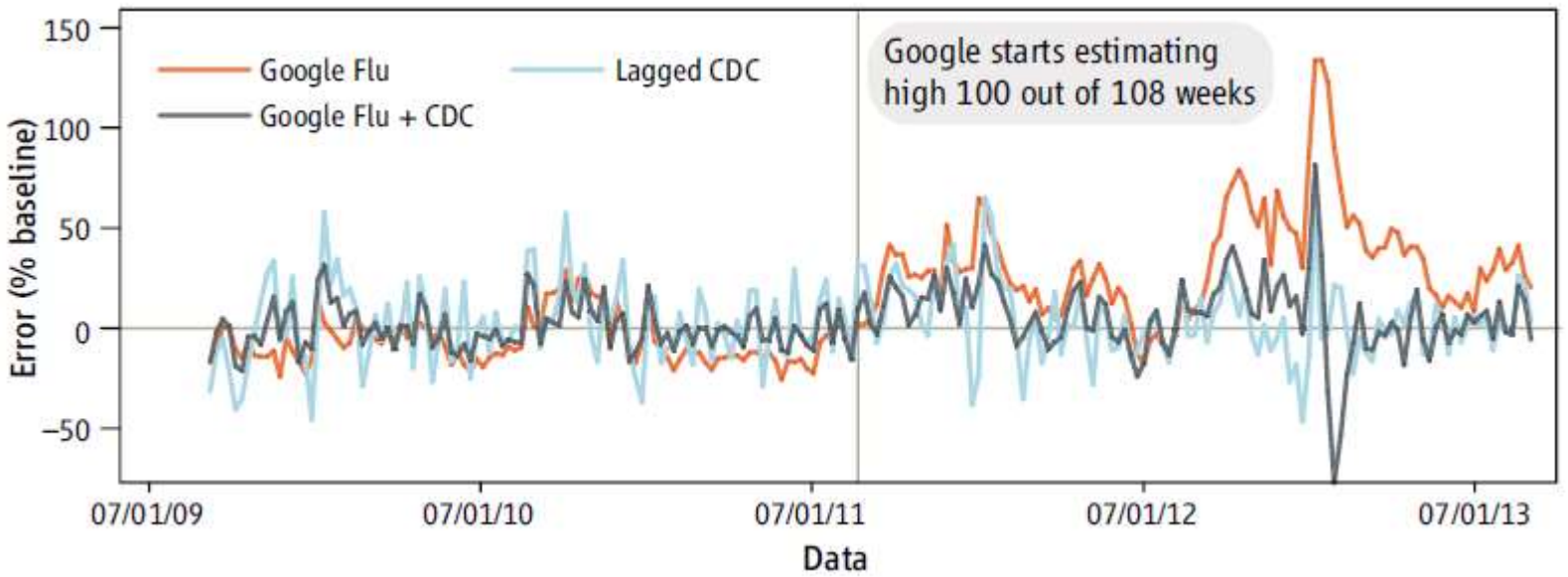
BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

David Lazer,[1,2*] Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[3,5,6]

www.sciencemag.org SCIENCE VOL 343 14 MARCH 2014

Overestimation by 50% in 2012-2013



7. Comprendre pour mieux prédire

- Confusion entre corrélation et causalité
- Difficile d'inférer la causalité à partir de données d'observations.
  - Effet d'un traitement: certains individus ont eu X=1, d'autres non

Inférence causale et raisonnement contrefactuel

- L'identité de base de décomposition du résultat d'un traitement:

$$
\begin{aligned}
&\text{Outcome for treated} - \text{Outcome for untreated} \\
&= [\text{Outcome for treated} - \text{Outcome for treated if not treated}] \\
&\quad + [\text{Outcome for treated if not treated} \\
&\qquad - \text{Outcome for untreated}] \\
&= \text{Impact of treatment on treated} + \text{selection bias.}
\end{aligned}
$$

- Partie contrefactuelle : « Outcome for treated if not treated"

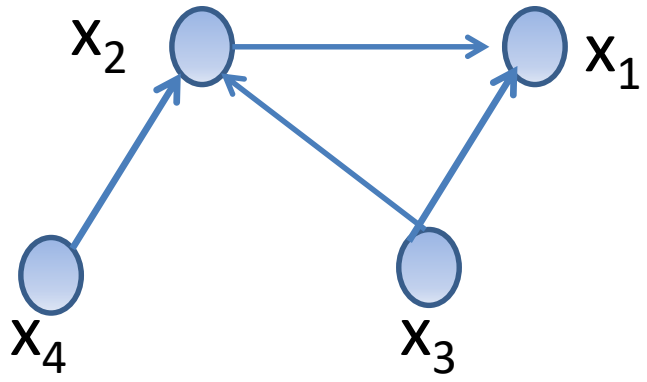
- Problème: on ne peut pas savoir ce qui serait arrivé aux individus traités s'ils ne l'avaient pas été!
- Estimer la partie contrefactuelle
  - Inférence causale de Rubin 1974
  - Propensity score matching de Rosenbaum et Rubin 1983
  - Pearl 2000

Le retour de l'expérimentation

- As Box et al. put it, "To find out what happens when you change something, it is necessary to change it." … the best way to answer causal questions is usually to run an experiment. (Varian, 2016)
- L'identité de base montre l'intérêt des essais randomisés: le biais de sélection est alors d'espérance nulle, d'où la possibilité d'estimer l'impact causal .
  - A/B testing, Marketing, publicité sur le web (Bottou,2013)

- Un modèle hybride: un schéma de régression (linéaire ou non) complété par un diagramme de causalité

$$\hat{y} = f(\mathbf{x})$$



DAG: Directed Acyclic Graph

Arthur M. Sackler COLLOQUIA

Drawing Causal Inference from Big Data

This meeting was held March 26-27, 2015 at the National Academy of Sciences 2101Constitution Ave. NW in Washington, D.C.
Organized by Richard M. Shiffrin (Indiana University), Susan Dumais (Microsoft Corporation), Mike Hawrylycz (Allen Institute), Jennifer Hill (New York University), Michael Jordan (University of California, Berkeley), Bernhard Schölkopf (Max Planck Institute) and Jasjeet Sekhon (University of California, Berkeley)
Graduate Student / Postdoctoral Researcher travel awards sponsored by the National Science Foundation and the Ford Foundation.

July 5, 2016
vol. 113 no. 27

http://www.nasonline.org/programs/sackler-colloquia/completed_colloquia/Big-data.html

https://www.pnas.org/big_data

PS Political Science & Politics — American Political Science Association

Symposium: Big Data, Causal Inference, and Formal Theory: Contradictory Trends in Political Science?

Big Data, Causal Inference, and Formal Theory: Contradictory Trends in Political Science? - Introduction

William Roberts Clark and Matt Golder

PS: Political Science & Politics / Volume 48 / Issue 01 / janvier 2015, pp 65 - 70

Can Big Data Solve the Fundamental Problem of Causal Inference?

Rocío Titiunik

PS: Political Science & Politics / Volume 48 / Issue 01 / janvier 2015, pp 75 - 79

En guise de conclusion

- Ambiguité du mot modèle
- Modèles pour comprendre (génératifs): la statistique comme auxiliaire de la science. Modèles pour prédire: l'autre face de la statistique, combinée avec les méthodes du Machine Learning: l'aide à la décision.
  - Un bon modèle prédictif améliore la connaissance
- Les modèles simples ne le sont pas tant que cela!
- Convergence entre inférence empirique et inférence causale

References

- C.Anderson (2008) The End of Theory: The Data Deluge Makes the Scientific Method Obsolete, http://www.wired.com/2008/06/pb-theory/
- L.Bottou et al. (2013) Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising, *Journal of Machine Learning Research*, 14, 3207–3260,
- Breiman, L., (1996): Stacked Regressions. *Machine Learning*, 24:49-64
- L.Breiman (2001) Statistical Modeling: The Two Cultures, *Statistical Science*, 16, 3, 199–231
- B.Clarke (2003) Comparing Bayes Model Averaging and Stacking When ModelApproximation Error Cannot be Ignored, *Journal of Machine Learning Research,* 4, 683-712
- D.Donoho (2015) 50 years of Data Science, *Tukey Centennial workshop*, https://dl.dropboxusercontent.com/u/23421017/50YearsDataScience.pdf

- H. Noçairi , C. Gomes , M. Thomas , G. Saporta (2016) Improving Stacking Methodology for Combining Classifiers; Applications to Cosmetic Industry, *Electronic Journal of Applied Statistical Analysis*, vol. 9(2), 340-361
- U.Grömping, (2015). Variable importance in regression models. *WIREs Computational Statistics,* 7, 137-152.
- G.Saporta (2008) Models for Understanding versus Models for Prediction, In P.Brito, ed., *Compstat Proceedings*, Physica Verlag, 315-322
- G. Shmueli (2010) To explain or to predict? *Statistical Science,* 25**,** 289–310
- V.Vapnik (2006) *Estimation of Dependences Based on Empirical Data*, 2nd edition, Springer
- H.Varian (2014) Big Data: New Tricks for Econometrics, *Journal of Economic Perspectives*, 28, 2, 3–28
- H.Varian (2016) Causal inference in economics and marketing, *PNAS* , 113, 7310-7315
- H.Wallard (2015) Using Explained Variance Allocation to analyse Importance of Predictors, *16th ASMDA conference proceedings*, 1043-1054